Nucleic Acids Research, 2020 **1**
doi: 10.1093/nar/gkaa1068

GreenPhylDB v5: a comparative pangenomic database for plant genomes

Guignon Valentin [1,2,*], Toure Abdel[3], Droc Gaëtan [2,4,5], Dufayard Jean-François[2,4,5], Conte Matthieu[3] and Rouard Mathieu [1,2,*]

[1]Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier, France, [2]French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, F-34398 Montpellier France, [3]Syngenta Seeds SAS, 31790 Saint-Sauveur France, [4]AGAP, Univ de Montpellier, CIRAD, INRAE, Montpellier SupAgro, F-34398 Montpellier, France and [5]CIRAD, UMR AGAP, F-34398 Montpellier, France

Received September 04, 2020; Revised October 19, 2020; Editorial Decision October 20, 2020; Accepted October 21, 2020

ABSTRACT

Comparative genomics is the analysis of genomic relationships among different species and serves as a significant base for evolutionary and functional genomic studies. GreenPhylDB (https://www.greenphyl.org) is a database designed to facilitate the exploration of gene families and homologous relationships among plant genomes, including staple crops critically important for global food security. GreenPhylDB is available since 2007, after the release of the *Arabidopsis thaliana* and *Oryza sativa* genomes and has undergone multiple releases. With the number of plant genomes currently available, it becomes challenging to select a single reference for comparative genomics studies but there is still a lack of databases taking advantage several genomes by species for orthology detection. GreenPhylDBv5 introduces the concept of comparative pangenomics by harnessing multiple genome sequences by species. We created 19 pangenes and processed them with other species still relying on one genome. In total, 46 plant species were considered to build gene families and predict their homologous relationships through phylogenetic-based analyses. In addition, since the previous publication, we rejuvenated the website and included a new set of original tools including protein-domain combination, tree topologies searches and a section for users to store their own results in order to support community curation efforts.

INTRODUCTION

Plant comparative genomics resources usually compare reference genomes to compute homology sequences and enable functional annotation transfer (1,2,3,4,5). However, with the growing number of whole genome sequences available within the same species, it has been shown than a single reference is not enough to capture its total genetic diversity (6). A pangenome, usually defined as the full gene repertoire within a species, can be partitioned into core genes that are shared by all individuals and dispensable genes that are present only in a subset of individuals (6,7,8). Characterizing them can have a great potential in plants for crop improvement (7,9,10) as candidate genes can potentially be missing in the genotype used to set up a reference genome. Pangenomic studies have recently been conducted in several crops, revealing significant differences with presence absence variations (PAVs) and/or copy number variations across genotypes (11,12,13,14,15,16,17,18). It became obvious that distinguishing core from dispensable genes is important as dispensable genes can be associated with useful trait diversity (10). Finally, PAVs can have an influence on orthology detection as specific genotype gene losses can lead to false negative results in interspecific comparisons or to pseudo-orthology (19).

Until now, comparative genomics databases have not fully taken advantage of these new datasets. Here, we present an updated version of the GreenPhylDB, a database that features multiple genomes for 19 species (e.g. rice, maize, banana, grape and cacao) as well as 27 other species with single reference genomes, for a total of 46 genomes. Publicly available genomes were processed to generate representative pangenes (i.e. a set of representative or consensus sequences) for species that were used in multi-species sequence clustering. Resulting gene families were functionally annotated and analysed with orthology detection methods.

*To whom correspondence should be addressed. Tel: +33 467612908; Fax: +33 467610334; Email: m.rouard@cgiar.org
Correspondence may also be addressed to Guignon Valentin. Email: v.guignon@cgiar.org

© The Author(s) 2020. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa1068/6006197 by guest on 11 December 2020

DATABASE CONSTRUCTION

Sequence retrieval and quality checks

We retrieved 132 publicly available datasets (coding DNA sequence and protein-coding genes) for 46 (Supplementary Table S1) and assessed their gene annotation predictions using BUSCO Plants v3.0.2 (embryophita_odb10) (20). We checked that the number of CDS was consistent with the number of proteins and that they shared the same locus tag name. When protein-coding genes were missing, we generated the sequences from GFF files originating from data providers. Finally, alternate splices were filtered, and the longest sequence was conserved.

Pangene construction

Out of the full dataset, 105 genomes were considered to produce 19 pangenes computed with the get_homologues-est software v20092018 (21) (Table 1), based on NCBI Blast-v2.2 and using the following program options (-M -F -t 0 -m cluster) as previously applied to the *Brachypodium distachyon* dataset (13). We processed each cluster the following way :

i. For single-gene copy clusters (a single sequence per genome), protein sequences were aligned using MAFFT v7.313 (22) (parameters adjusted according to the number of sequences) and an automatic procedure to generate a consensus sequence was applied (Figure 1A). For each position of the alignment, we kept the most frequent amino acid. In case of a tie, the amino acid of the genome with the highest BUSCO scores ('complete' then 'fragmented' and finally lowest 'missing' scores) was selected. Finally, if more gaps than amino acids were present, this position was removed from the sequence.
ii. For multi-copy clusters (multiple sequences per genome), we applied the same procedure as for single-gene copy clusters but added a preliminary step to select a representative sequence by cluster. Multiple sequence alignments were used to generate a distance matrix using distmat (Jukes-Cantor correction method) from EMBOSS v6.6 (Figure 1B). The matrix was required to define the distance for each sequence of all other genomes and the sequences with the smallest sum of distance was selected as representative of the considered genome (Figure 1C). Then, the consensus step was applied. It is worth mentioning that get_homologs-est generated sequence clusters of not too distantly related sequences. Large gene families can include several multi-copy clusters being grouped together at the sequence clustering step.
iii. For genotype-specific clusters (paralogs in a single genome), we generated a distance matrix between all sequences and the sequence with the lowest average distance (min(d/sum(d))) of all sequences was putatively considered as the most representative sequence. Those sequences were added to the pangene.
iv. Finally, singletons (cluster of one sequence) were searched for similarity using DIAMOND (23) with a default *e*-value on the protein-coding genes of all other species genomes to predict their putative prediction accuracy. Sequences with a minimum of one hit in at least two species were added to the pangene; otherwise sequences were excluded.

As a unique identifier was required for each pangene, we defined a nomenclature with a prefix composed of the [5-letter UniProt taxonomy database code]_pan, followed by p (for protein) and an auto-increment of 6-digits (e.g. musac_pan_p029014 for *Musa acuminata*).

Sequence clustering and functional annotation

Pangenes and protein-coding genes of reference genomes (without pangenes) were searched all against all using DIAMOND. We then performed a clustering using TribeMCL (24) ($M = 1.2$, 2, 3 and 5), defining 4 levels of stringency (from 1 to 4) to take into account potential subclassification and we obtained 9419, 18 805, 23 409 and 29 345 clusters, respectively.

We then scanned all sequences for protein domain signatures using InterProscan (25,26) and also crossed linked matches with UniProtKB-SwissProt entries (27). Cluster names resulting from curation from previous GreenPhylDB versions (2,28) were transferred when at least 51% of sequences were found clustered together as before (based on species in common between releases). In addition, for this release, we implemented an automatic method to name clusters based on the name of InterPro domains (family type only) that were found specific to clusters. In other words, when detected in at least 51% of the sequences composing an unannotated cluster, the name of the InterPro signature was assigned to it. In total, GreenPhylDB comprises 3538 clusters functionally characterized across the four levels.

Homology inference

The previous phylogenetic-based methodology that we applied in the previous version has been conserved but uses a larger set of genomes to update our automated pipeline. The pipeline uses MAFFT for the multiple alignment step. FastTree 2 (v2.1.11) (29) was preferred over PhyML (30) due to the size of the clusters. Gene rooting and orthologous scoring was computed with Rap-Green (31) using the *viridiplantae* species tree extracted from NCBI taxonomy and converted into PhyloXML (2). We successfully produced gene trees at level 1 for more than 99.8% of the clusters ($n = 9413$) which enabled us to predict ∼17.8 million of orthologs and ∼1.8 million of in-paralogs (or ultra-paralogs) relationships. The pipeline was complemented by a Reciprocal Best Hits (RBH) method—computed between all pairs of genomes—that resulted in more than ∼12.1 million orthologous relationships.

USING GREENPHYL

With this updated version, the website has received a face-lift. It now takes advantage of the bootstrap and D3.js frameworks to improve the user experience and to be more responsive. Alternatively, it can also be accessed programmatically using Resource Description Framework (RDF) as

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa1068/6006197 by guest on 11 December 2020

Table 1. List of GreenPhylDB pangenes with associated statistics

Species Code	Genomes available by species name	# Genes	Busco complete (%)	# Pangenes	% singletons (not in pangene)
BRADI	*Brachypodium distachyon (54 genomes)*	44 858 (average)	95.5 (average)	61 622	0.8
BRANA	*Brassica napus*	101 040	98.3	77 456	5.0
		80 382	96.2		0.5
		70 162	91.2		0.6
BRAOL	*Brassicaoleracea*	35 400	80.7	60 869	2.4
		61 279	96.9		7.7
		56 687	99.5		3.3
BRARR	*Brassica rapa*	46 250	98.1	49 916	4.4
		46 721	97		4.1
CAPAN	*Capsicum annuum*	35 336	90.6	41 828	7.4
		34 476	84.6		6.0
		35 884	88.9		8.2
CICAR	*Cicer arietinum*	28 269	94.8	25 013	7.5
		30 257	93		11.7
COCNU	*Cocos nucifiera*	52 931	87.8	38 584	1.4
		34 953	86.3		7.0
CUCSA	*Cucumis sativus*	22 324	89.5	23 446	11.2
		23 780	94.8		7.5
		22 935	94.8		5.7
IPOTF	*Ipomoea trifida*	32 301	96.6	21 417	8.0
		30 227	94.7		5.3
MAIZE	*Zea mays*	39 591	94.6	45 301	4.6
		40 003	92.6		6.7
		40 557	87.3		9.4
		36 509	87.8		4.9
MALDO	*Malus domestica*	45 116	98.3	54 987	7.5
		95 232	91		24.5
		44 677	95.2		15.8
MEDTR	*Medicago truncatula*	50 444	96.6	43 859	19.2
		44 623	98.8		4.7
MUSAC	*Musa acuminata*	35 276	98.5	45905	5.5
		44 702	60.3		17.5
		32 692	71.2		18.2
		45 069	71.9		21.0
ORYSA	*Oryza sativa*	55 986	95.1	56785	11.5
		36 140	87.8		11.3
		37 549	96.1		11.1
		60 897	93.7		9.9
		60 123	89.4		10.7
		35 495	89.9		5.1
		35 594	99.3		2.3
SORBI	*Sorghum bicolor*	34 129	99.2	45054	11.8
		36 110	97.4		15.4
SOYBN	*Glycine max*	54 175	99.5	34512	5.0
		52 130	99.3		4.1
THECC	*Theobroma cacao*	21 330	99.2	32917	1.9
		44 607	99.6		23.1
TRITU	*Triticum turgidum*	107 891	99.6	54687	8.6
		67 182	98.9		5.3
VITVI	*Vitis vinifera*	41 733	98.4	43766	22.8
		96 331	92.4		5.1
		73 109	95.9		5.6

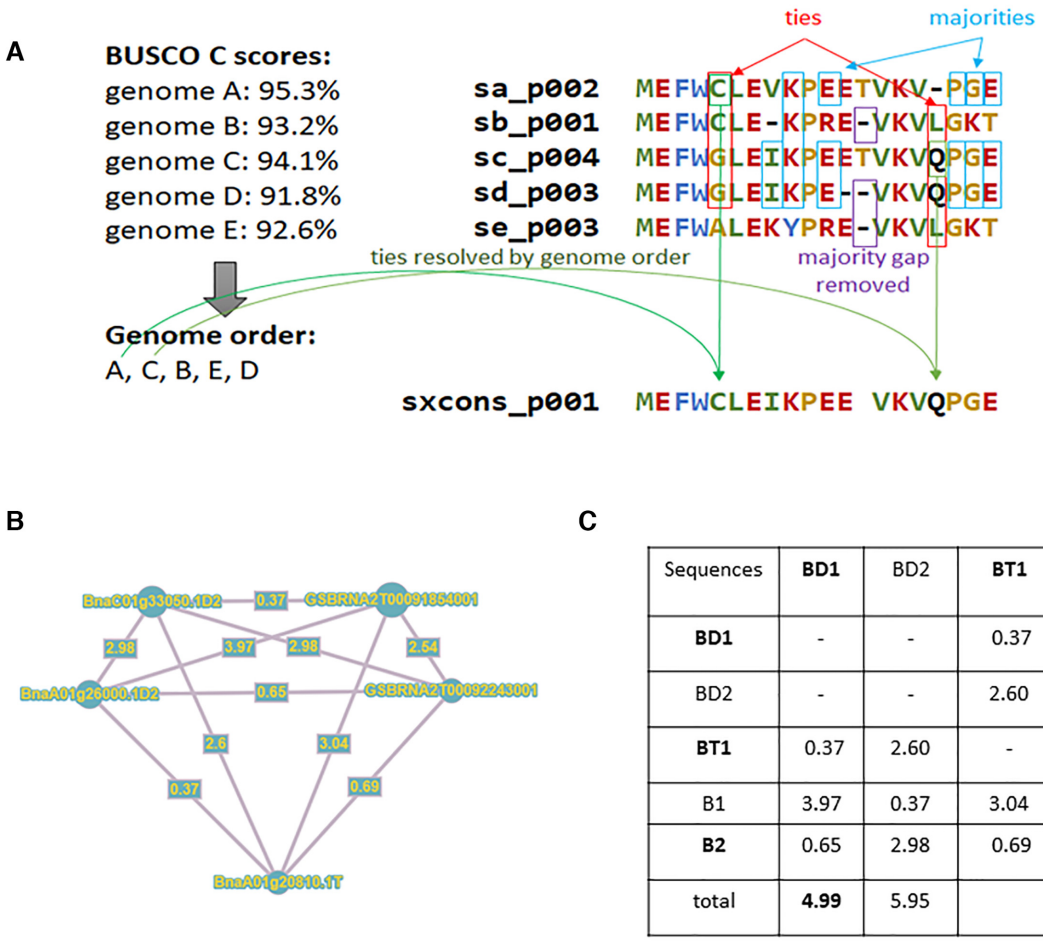
implemented in AgroLD, a knowledge-based system relying on semantic web technologies (32).

Gene family pages

All cluster (or gene family) pages present the same type of information divided into several tabs :

1. *Gene family composition: a* bar chart allows users to visualize at a glance the composition of the gene family by species (Figure 2A). Species are ordered taxonomically to easily detect possible variations between phyla. Each bar is clickable and produces a table with the list of sequences and associated cross-references (i.e. InterPro, UniProt). Sequences can be exported in multiple formats and/or stored in a user list.
2. *Gene family structure:* sequences are clustered at four levels of clustering, from less stringent to more stringent, in most cases narrowing the number of sequences (Figure 2B).
3. *Protein domains:* here, InterProscan was used to assess the domain conservation consistency and the specificity of the sequence clusters (Figure 2C). For each cluster, we performed statistical analyses to determine whether InterPro signatures were specific and therefore

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa1068/6006197 by guest on 11 December 2020



Sequences	**BD1**	BD2	**BT1**	B1	**B2**
BD1	-	-	0.37	3.97	0.65
BD2	-	-	2.60	0.37	2.98
BT1	0.37	2.60	-	3.04	0.69
B1	3.97	0.37	3.04	-	-
B2	0.65	2.98	0.69	-	-
total	**4.99**	5.95		7.38	**4.32**

Figure 1. Pangene construction. (**A**) Schema illustrating the creation of consensus sequences. (**B**) Example of distance matrix network with five sequences from three genomes of *Brassica napus* (brana_pan_p029014). (**C**) Selection of representative sequences based on the minimum value of summed genetic distances. Sequences conserved are in bold.

not found in any other sequences of clusters of the same level.

4. *Phylogenomic analyses:* this section includes multiple sequence alignments that can be downloaded and visualised using MSAviewer (33) and gene trees can be explored with InTreeGreat (https://www.southgreen.fr/content/intreegreat-tool) and PhyD3 (34) (Figure 2D). Some gene trees can be very large, and the interface proposes an option to prune automatically the tree based on user choice for a range of species.
5. *Homologous predictions:* the interface enables users to display and refine all the homologies detected by the phylogenetic-based approach and Reciprocal Best Hits (RBH). It is possible to filter and select only a subset of species of interest.

Pangene pages

The new page type for pangene sequences is a central and unique concept in this version as they were used for the clustering and homology predictions instead of all individual sequences that compose it. When browsing these pages, users can quickly see which genes are present or missing by looking at the status: core or dispensable compartments. Then, information related to the sequence composition is reported. Users can access information about pangene classification, the consensus sequence (except for singletons) with the multiple sequence alignments used to create it as well as related homology predictions (Figure 3). In the case of multi-copy clusters, it is possible to see which sequence was selected as representative (.rep) or participant (.p) and also why they were selected by browsing the distance matrix.

New tools

The database can still be searched via keyword searches or by entering a query sequence for similarity search using DIAMOND (which replaces BLAST for faster processing), enhanced by new tools to further explore those datasets.

Quick search. A new interface has been designed to retrieve in a comprehensive and concise way all the information associated with gene family names, sequence annotation and annotations from InterPro and UniProt mappings.

IPR2genomes—InterPro domain search. It is now possible to search sequences and associated clusters based on a combination of InterPro domain signatures. This can be partic-

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa1068/6006197 by guest on 11 December 2020

Figure 2. Overview of gene family interfaces for GP001047: U2 auxiliary factor small subunit family. (**A**) diagram of the sequence count (327) by species (46) (**B**) Sequence flow in cluster structure (from level 1 to 4). (**C**) InterPro domain specificity statistics. About 92% of sequences have the U2 auxiliary factor small subunit which is uniquely found in this cluster. (**D**) Viewers for multiple sequence alignment (MSAViewer) and phylogenetic tree (InTreeGreat).

ularly helpful when searching for transcription factors for which sequences must contain some domains but not others (35) as the Markov Cluster Algorithm (MCL) may fail grouping them accurately. The interface allows the use of various operators (e.g. AND, OR, NOT, ONLY) to filter a set of sequences for all genomes. Results can be compared with the MCL automatic clustering to check consistencies or differences.

TreePattern —Tree topology search. A tree search can be done by filtering on gene tree topologies (31). Users can draw the topology with species or taxonomic groups as leaves or nodes of the tree and apply constraints such presence or absence of duplications. Resulting trees can be accessed individually (or exported in bulk results as CSV file) and defined patterns are highlighted. This feature is useful for identifying gene families with an expected evolutionary scenario due to gene duplications.

Manual curation and sharing of gene families

While automatic clustering is a relevant and efficient starting point, sometimes limitations (e.g. missing sequences, errors in gene annotations) are present and prevent access to ready-to-use datasets, justifying a deeper characterization that will eventually lead to a refinement of the automatic clustering. As a result, knowledge generated on individual gene families is often available only in publications and their supplementary information as PDFs. To encourage knowledge capture, we developed a section for advanced users to create and share their own gene families. Two methods are possible: users can either start from scratch and upload their data or use existing clusters and take advantage of multiples operations implemented in the 'MyList' features: such feature was indeed developed to facilitate intersecting, combining clusters. This new tool can be valuable during the review process by providing a unique identifier to referees—and eventually to users—to explore the structure and composition of the submitted gene family.

USE CASES

In this section, we describe three possible uses that are enabled by this new GreenPhylDB version. Concrete examples related to each of them are further documented in Supplementary Data.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa1068/6006197 by guest on 11 December 2020

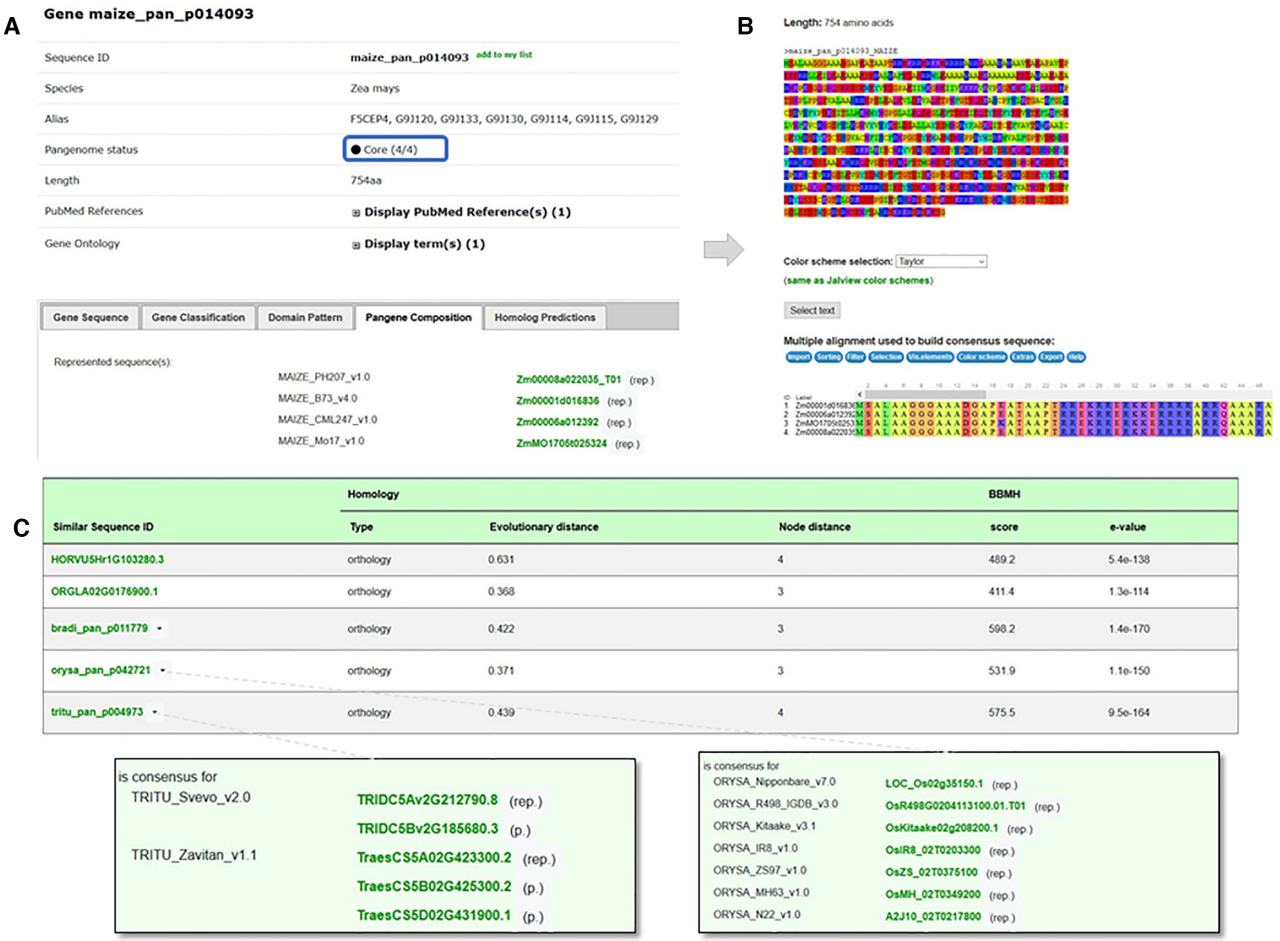


Figure 3. Example of a pangene page (i.e. maize_pan_p014093) member of the U2 auxiliary factor small subunit gene family (GP001047). (**A**) Gene composition tab: all genes are part of core compartment for those four genomes since a representative sequence exists for each of the reference genomes. (**B**) Consensus sequences and associated multiple alignment. (**C**) List of homologs: the *Zea mays* pangene is predicted orthologous to pangenes in *Triticum turgidum* and in *Oryza sativa*. The Popups (green rectangles) display the sequence compositions of the respective pangenes. (.rep) refers to the representative sequence kept to create the consensus and (.p) to paralog sequences not used in the consensus.

i. You want to analyse a gene family with focus on a specific species sampling.
   a. search the gene family by keyword(s), locus gene ids or sequences using dedicated search families (toolbox menu).
   b. browse the family structure and explore sub clusters at level 2, 3 or 4.
   c. browse the family composition of the cluster (or a specific sub cluster) to list all the sequences and pangenes selecting one or several bars in the diagram. Export the selected sequences of interest using proposed file formats.
   d. (optional) if the gene family sequences are characterized by several protein domains, check possible additional sequences in the database using IPR2genomes (toolbox menu). For individual protein domains, its specificity is indicated in each gene family page (if specific, no need to search in other gene families).
   e. (optional) In case of additional analyses (e.g. addition of sequences from a new sequences genomes), you can create a 'custom family' by uploading the sequences on the website and share the link in a manuscript, with collaborators or reviewers during the review process for a user-friendly exploration of the dataset.

ii. You are interested in finding genes that are linked to a specific evolutionary scenario (e.g. duplicated genes in one species but not in another)
   a. retrieve full list of gene trees and related gene families using TreePattern (toolbox menu).
   b. browse examples gene families to see patterns (highlighted with dashed lines in the tree).
   c. click on the sequence name to access the family.
   d. browse the family composition of the cluster (or a specific sub cluster) to list all the sequences and pangenes selecting one or several bars in the diagram. Export the selected sequences of interest using proposed file formats

iii. You have a candidate gene in rice, maize or banana (or any of the 19 pangenes) and want to retrieve the related sequences in other genomes of the same species

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa1068/6006197 by guest on 11 December 2020

and then find orthologs in other species (e.g. *Arabidopsis*).

a. search by sequence or by locus ID to identify the pangene ID.
b. retrieve the pangene composition to get all members and check the status (core or dispensable).
c. (optional) check the multiple gene alignment to see level of divergence.
d. go to the gene family and explore (or download) the gene tree.
e. retrieve predicted orthologs (by phylogeny and/or Reciprocal Best Hits). Alternatively, use the homologous sequence search directly (toolbox menu).

CONCLUSION

This new version of GreenPhylDB provides a unique way to scale up plant comparative genomics studies across multiple plants species by leveraging pangenomic datasets. This release paves the way to the transition from reference-based genomics to pangenome-based systems and tools. In this context, the website includes new powerful search interfaces to explore the content of the gene family collection. Advanced users can also deposit the results of their expert gene family curation for further use and reference. GreenPhylDB is an important resource to understand the genetic basis of genome diversity among plant species and has the potential to accelerate gene discovery to support crop improvement.

DATA AVAILABILITY

All datasets produced by our automatic analyses are accessible via GreenPhylDB user interfaces or can be downloaded at https://www.greenphyl.org/cgi-bin/downloads.cgi.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was technically supported by the CIRAD–UMR AGAP HPC Data Center of the South Green Bioinformatics platform (https://www.southgreen.fr/). We thank the scientists who facilitated access to some genome datasets: Stéphanie Bocs (CIRAD) and Darlon Lantican (UPLB) for coconut, Ming Li (SAAS) for wild sweet potato, Stéphane Deschamps and Helen Lin (Corteva) for sorghum. We also acknowledge community databases which provided access to the genome datasets, thus facilitating the development of this online resource.

FUNDING

Syngenta Seeds SAS; CGIAR Research Program, Roots, Tubers and Bananas. Funding for open access charge: Syngenta Seeds SAS.
Conflict of interest statement. None declared.

REFERENCES

1. Van Bel,M., Diels,T., Vancaester,E., Kreft,L., Botzki,A., Van de Peer,Y., Coppens,F. and Vandepoele,K. (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.*, **46**, D1190–D1196.
2. Rouard,M., Guignon,V., Aluome,C., Laporte,M.-A., Droc,G., Walde,C., Zmasek,C.M., Périn,C. and Conte,M.G. (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res.*, **39**, D1095–D1102.
3. Goodstein,D.M., Shu,S., Howson,R., Neupane,R., Hayes,R.D., Fazo,J., Mitros,T., Dirks,W., Hellsten,U., Putnam,N. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.
4. Gupta,P., Naithani,S., Tello-Ruiz,M.K., Chougule,K., D'Eustachio,P., Fabregat,A., Jiao,Y., Keays,M., Lee,Y.K., Kumari,S. *et al.* (2016) Gramene database: navigating plant comparative genomics resources. *Curr. Plant Biol.*, **7–8**, 10–15.
5. Bolser,D., Staines,D.M., Pritchard,E. and Kersey,P. (2016) Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. In: Edwards,D. (ed). *Plant Bioinformatics: Methods and Protocols, Methods in Molecular Biology*. Springer, NY, pp. 115–140.
6. Golicz,A.A., Batley,J. and Edwards,D. (2016) Towards plant pangenomics. *Plant Biotechnol. J.*, **14**, 1099–1105.
7. Tranchant-Dubreuil,C., Rouard,M. and Sabot,F. (2019) Plant pangenome: impacts on phenotypes and evolution. In: Roberts,J.A. (ed). *Annual Plant Reviews Online*. pp. 453–478.
8. Marschall,T., Marz,M., Abeel,T., Dijkstra,L., Dutilh,B.E., Ghaffaari,A., Kersey,P., Kloosterman,W.P., Mäkinen,V., Novak,A.M. *et al.* (2018) Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.*, **19**, 118–135.
9. Tao,Y., Zhao,X., Mace,E., Henry,R. and Jordan,D. (2019) Exploring and exploiting pan-genomics for crop improvement. *Mol. Plant*, **12**, 156–169.
10. Gabur,I., Chawla,H.S., Snowdon,R.J. and Parkin,I.A.P. (2019) Connecting genome structural variation with complex traits in crop plants. *Theor. Appl. Genet.*, **132**, 733–750.
11. Golicz,A.A., Bayer,P.E., Barker,G.C., Edger,P.P., Kim,H., Martinez,P.A., Chan,C.K.K., Severn-Ellis,A., McCombie,W.R., Parkin,I.A.P. *et al.* (2016) The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.*, **7**, 13390.
12. Montenegro,J.D., Golicz,A.A., Bayer,P.E., Hurgobin,B., Lee,H., Chan,C.-K.K., Visendi,P., Lai,K., Doležel,J., Batley,J. *et al.* (2017) The pangenome of hexaploid bread wheat. *Plant J.*, **90**, 1007–1013.
13. Gordon,S.P., Contreras-Moreira,B., Woods,D.P., Marais,D.L.D., Burgess,D., Shu,S., Stritt,C., Roulin,A.C., Schackwitz,W., Tyler,L. *et al.* (2017) Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.*, **8**, 2184.
14. Wang,W., Mauleon,R., Hu,Z., Chebotarov,D., Tai,S., Wu,Z., Li,M., Zheng,T., Fuentes,R.R., Zhang,F. *et al.* (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.
15. Zhao,Q., Feng,Q., Lu,H., Li,Y., Wang,A., Tian,Q., Zhan,Q., Lu,Y., Zhang,L., Huang,T. *et al.* (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.*, **50**, 278–284.
16. Hirsch,C.N., Foerster,J.M., Johnson,J.M., Sekhon,R.S., Muttoni,G., Vaillancourt,B., Peñagaricano,F., Lindquist,E., Pedraza,M.A., Barry,K. *et al.* (2014) Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*, **26**, 121–135.
17. Gao,L., Gonda,I., Sun,H., Ma,Q., Bao,K., Tieman,D.M., Burzynski-Chang,E.A., Fish,T.L., Stromberg,K.A., Sacks,G.L. *et al.* (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.*, **51**, 1044–1051.
18. Hübner,S., Bercovich,N., Todesco,M., Mandel,J.R., Odenheimer,J., Ziegler,E., Lee,J.S., Baute,G.J., Owens,G.L., Grassa,C.J. *et al.* (2019) Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants*, **5**, 54–62.
19. Koonin,E.V. (2005) Orthologs, paralogs and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
20. Simão,F.A., Waterhouse,R.M., Ioannidis,P., Kriventseva,E.V. and Zdobnov,E.M. (2015) BUSCO: assessing genome assembly and

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa1068/6006197 by guest on 11 December 2020

annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
21. Contreras-Moreira,B., Cantalapiedra,C.P., García-Pereira,M.J., Gordon,S.P., Vogel,J.P., Igartua,E., Casas,A.M. and Vinuesa,P. (2017) Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. *Front. Plant Sci.*, **8**, 184.
22. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
23. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
24. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
25. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
26. Mitchell,A.L., Attwood,T.K., Babbitt,P.C., Blum,M., Bork,P., Bridge,A., Brown,S.D., Chang,H.-Y., El-Gebali,S., Fraser,M.I. *et al.* (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.*, **47**, D351–D360.
27. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*, **2011**, bar009.
28. Conte,M.G., Gaillard,S., Lanau,N., Rouard,M. and Périn,C. (2008) GreenPhylDB: a database for plant comparative genomics. *Nucleic Acids Res.*, **36**, D991–D998.
29. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
30. Guindon,S., Delsuc,F., Dufayard,J.-F. and Gascuel,O. (2009) Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.*, **537**, 113–137.
31. Dufayard,J.-F., Duret,L., Penel,S., Gouy,M., Rechenmann,F. and Perrière,G. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, **21**, 2596–2603.
32. Venkatesan,A., Ngompe,G.T., Hassouni,N.E., Chentli,I., Guignon,V., Jonquet,C., Ruiz,M. and Larmande,P. (2018) Agronomic Linked Data (AgroLD): a knowledge-based system to enable integrative biology in agronomy. *PLoS One*, **13**, e0198270.
33. Yachdav,G., Wilzbach,S., Rauscher,B., Sheridan,R., Sillitoe,I., Procter,J., Lewis,S.E., Rost,B. and Goldberg,T. (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, **32**, 3501–3503.
34. Kreft,L., Botzki,A., Coppens,F., Vandepoele,K. and Van Bel,M. (2017) PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics*, **33**, 2946–2947.
35. Lang,D., Weiche,B., Timmerhaus,G., Richardt,S., Riano-Pachon,D.M., Correa,L.G.G., Reski,R., Mueller-Roeber,B. and Rensing,S.A. (2010) Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome. Biol. Evol.*, **2**, 488–503.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa1068/6006197 by guest on 11 December 2020